PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-132811
(43)Date of publication of application : 10.05.2002

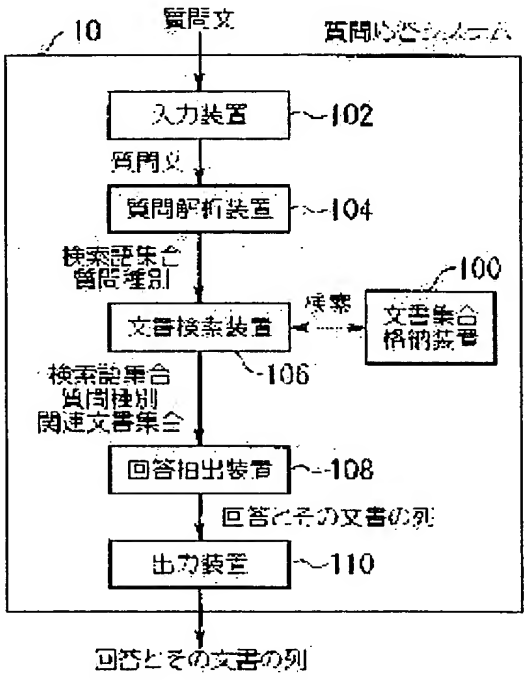
(51)Int.Cl. G06F 17/30
G06F 17/27
G06F 17/28

(21)Application number : 2000-319998
(22)Date of filing : 19.10.2000

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : SASAKI YUTAKA
ISOZAKI HIDEKI
TAIRA HIROYORI
KAZAWA HIDETO
HIROTA KEIICHI
NAKAJIMA HIROYUKI
HIRAO TSUTOMU
KATO TSUNEAKI

(54) METHOD AND SYSTEM FOR ANSWERING QUESTION AND RECORDING MEDIUM WITH RECORDED QUESTION ANSWERING PROGRAM

(57)Abstract:
PROBLEM TO BE SOLVED: To output an answer to a question sentence and its document when a document collection is given.
SOLUTION: This question answering system outputs the answer to the question sentence and its document once given the document collection and question sentence. The system has a document collection storage device 100 which stores document collections, an input device 102 which receives the question sentence, a question analyzing device 104 which decides a retrieval word collection and the kind of the question from the question sentence obtained from an input device, a document retrieving device 106 which retrieves a related document collection from the document collections stored in the document collection storage devices, an answer extracting device 108 which extracts answers to the question sentence from respective documents in the related document collection and generates an array of the answers and the documents having the answers extracted as an answer result to the question sentence, and an output device 110 which outputs the answer result.



LEGAL STATUS

[Date of request for examination] 21.12.2001

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than

the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－132811
(P2002－132811A)
(43)公開日　平成14年５月10日(2002.5.10)

(51)Int.Cl.'	識別記号	FI		テーマコード'(参考)
G06F 17/30	330	G06F 17/30	330C	5B075
	170		170A	5B091
17/27		17/27	E	
17/28		17/28	T	

審査請求 有　請求項の数9 OL （全 10 頁）

(21)出願番号　特願2000－319998(P2000－319998)

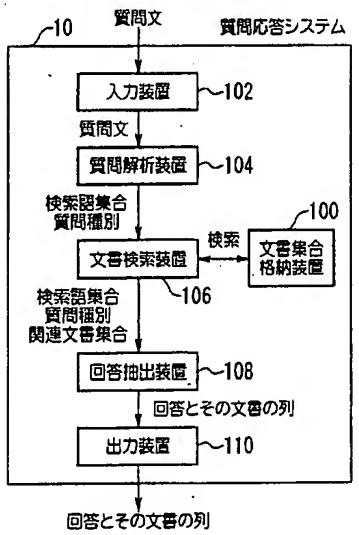
(22)出願日　平成12年10月19日(2000.10.19)

(71)出願人 000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者 佐々木 裕
東京都千代田区大手町二丁目３番１号 日本電信電話株式会社内
(72)発明者 磯崎 秀樹
東京都千代田区大手町二丁目３番１号 日本電信電話株式会社内
(74)代理人 100064908
弁理士 志賀 正武

最終頁に続く

(54)【発明の名称】 質問応答方法、質問応答システム及び質問応答プログラムを記録した記録媒体

(57)【要約】
【課題】 文書集合が与えられると質問文に対する回答とその文書を出力できるようにする。
【解決手段】 文書集合と質問文が与えられると、質問文に対する回答と文書の列を出力する質問応答システムであって、文書集合を格納する文書集合格納装置１００と、質問文を受けとる入力装置１０２と、入力装置から得られた質問文から検索語集合と質問種別を判定する質問解析装置１０４と、前記検索語集合と質問種別に従って、文書集合格納装置に格納された文書集合から関連文書集合を検索する文書検索装置１０６と、関連文書集合中の各文書から質問文に対する回答を抽出し、該回答と該回答を抽出した文書の列を質問文に対する応答結果として作成する回答抽出装置１０８と、応答結果を出力する出力装置１１０とを有する。



【特許請求の範囲】
【請求項１】 文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力するコンピュータシステムを使用した質問応答方法であって、
入力された質問文から検索語集合と質問種別を判定し、該検索語集合および該質問種別に従って、前記与えられた文書集合から関連文書集合を検索し、該関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力することを特徴とする質問応答方法。
【請求項２】 前記関連文書集合の要素は文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索することを特徴とする請求項１に記載の質問応答方法。
【請求項３】 前記関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの２つのスコアに従って、回答と文書の列を順序付けることを特徴とする請求項１または２のいずれかに記載の質問応答方法。
【請求項４】 前記関連文書集の各文書から回答を抽出する際に、固有名詞や数値表現の認識を行なうことを特徴とする請求項１乃至３のいずれかに記載の質問応答方法。
【請求項５】 文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問応答システムであって、
文書集合を格納する文書集合格納装置と、
質問文を受けとる入力装置と、
該入力装置から得られた前記質問文から検索語集合と質問種別を判定する質問解析装置と、
前記検索語集合と質問種別に従って、前記文書集合格納装置に格納された文書集合から関連文書集合を検索する文書検索装置と、
関連文書集合中の各文書から質問文に対する回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として作成する回答抽出装置と、
前記応答結果を出力する出力装置と、
を有することを特徴とする質問応答システム。
【請求項６】 前記文書検索装置は、関連文書集合の要素を文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索することを特徴とする請求項５に記載の質問応答システム。
【請求項７】 前記回答抽出装置は、前記文書検索装置が関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記回答抽出装置が前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの２つのスコアに従って、回答と文書の列を順序付けることを特徴とする請求項５または６のいずれかに記載の質問応答システム。
【請求項８】 前記回答抽出装置は、前記関連文書集合の各文書から回答を抽出する際に、固有名詞や数値表現の認識を行なうことを特徴とする請求項５乃至７のいずれかに記載の質問応答システム。
【請求項９】 文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問応答を行うための質問応答プログラムを記録したコンピュータ読み取り可能な記録媒体において、
質問文を受け取る第１のステップと、
入力された質問文から検索語集合と質問種別を判定する第２のステップと、
前記検索語集合および質問種別に従って、前記与えられた文書集合から関連文書集合を検索する第３のステップと、
前記関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を作成する第４のステップと、
前記回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力する第５のステップと、
をコンピュータに実行させる質問応答プログラムを記録した記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、自然言語処理システム、言語処理システム、知識処理システム、情報検索システム、情報抽出システム等に用いられ、質問文に対する回答を出力する質問応答方法、質問応答システム及び質問応答プログラムを記録した記録媒体に関する。
【０００２】
【従来の技術】従来の情報検索技術は、与えられた文書集合から、ユーザの入力した質問に合致する文書の集合を取り出すものであった。また、従来の情報抽出技術は、与えられた１つの文書について、分野ごとに予め決められた項目を抽出するものであった。従来の日本語質問応答システムは、回答として単語や単語の列を出力していたが、その回答の元となる記事を同時に出力していなかった。
【０００３】さらに、固有名詞の抽出技術の利用や、質問文に対する回答を含む可能性のある文書を検索する際の文書スコア、及び回答を抽出する際の抽出スコアの両方スコアを総合的に用いて、質問に対して出力する回答と文書の列の順序を決める等の、回答の精度を向上させるための工夫が行なわれていなかった。
【０００４】
【発明が解決しようとする課題】まず、従来の情報検索技術は質問に対する結果を文書の単位で応答するため、ユーザが文書の内容を読まなければならない。例えば、「日本の首相は誰ですが？」という質問に対して、返ってきた結果に含まれる文書を読むことにより初めて、「森首相」という答が判る。文書を読まなければならな

いことは「森首相」といった直接的な回答が欲しいユーザにとっては非常に煩わしいという問題が有った。

【０００５】次に、従来の情報抽出技術は、予め決められた特定の分野でしか使えないため、任意の質問文に対する回答を出力する質問応答システムでは使えない。従来の日本語質問応答システムは、回答として単語や単語の列を出力していたが、回答の元となる記事を返していなかったため、回答を裏付ける情報をユーザが得ることができなかった。例えば、「ジンバブエの大統領は誰ですか?」という質問に対して、「ムガベ大統領」と回答が返ってきても、ユーザが本当に「ムガベ大統領」が正しいかどうかを確認できないという問題が有った。

【０００６】本発明はこのような事情に鑑みてなされたものであり、質問文に対する直接的な回答とその回答の元となる文書を提示することにより回答の信頼性をユーザが確認することができる質問応答方法、質問応答システム及び質問応答プログラムを記録した記録媒体を提供することを目的とする。

【０００７】

【課題を解決するための手段】上記目的を達成するために、請求項１に記載の発明は、文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力するコンピュータシステムを使用した質問応答方法であって、入力された質問文から検索語集合と質問種別を判定し、該検索語集合および該質問種別に従って、前記与えられた文書集合から関連文書集合を検索し、該関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力することを特徴とする。

【０００８】請求項１に記載の発明によれば、入力された質問文から検索語集合と質問種別を判定し、該検索語集合および該質問種別に従って、前記与えられた文書集合から関連文書集合を検索し、該関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力するようにしたので、質問文に対する直接的な回答とその回答の元となる文書を提示することにより回答の信頼性をユーザが確認することができる。

【０００９】また請求項２に記載の発明は、請求項１に記載の質問応答方法において、前記関連文書集合の要素は文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索することを特徴とする。

【００１０】請求項２に記載の発明によれば、請求項１に記載の質問応答方法において、前記関連文書集合の要素は文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索するようにしたので、直接的な回答が得られる。

【００１１】また、請求項３に記載の発明は、請求項１または２のいずれかに記載の質問応答方法において、前記関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの２つのスコアに従って、回答と文書の列を順序付けることを特徴とする。

【００１２】請求項３に記載の発明によれば、請求項１または２のいずれかに記載の質問応答方法において、前記関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの２つのスコアに従って、回答と文書の列を順序付けるようにしたので、質問文に対する回答の精度の向上が図れる。

【００１３】また、請求項４に記載の発明は、請求項１乃至３のいずれかに記載の質問応答方法において、前記関連文書集の各文書から回答を抽出する際に、固有名詞や数値表現の認識を行なうことを特徴とする。

【００１４】請求項４に記載の発明によれば、請求項１乃至３のいずれかに記載の質問応答方法において、前記関連文書集の各文書から回答を抽出する際に、固有名詞や数値表現の認識を行なうようにしたので、質問文に対する精度の向上が図れる。

【００１５】また、請求項５に記載の発明は、文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問応答システムであって、文書集合を格納する文書集合格納装置と、質問文を受けとる入力装置と、該入力装置から得られた前記質問文から検索語集合と質問種別を判定する質問解析装置と、前記検索語集合と質問種別に従って、前記文書集合格納装置に格納された文書集合から関連文書集合を検索する文書検索装置と、関連文書集合中の各文書から質問文に対する回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として作成する回答抽出装置と、前記応答結果を出力する出力装置とを有することを特徴とする。

【００１６】請求項５に記載の発明によれば、文書集合を格納する文書集合格納装置と、質問文を受けとる入力装置と、該入力装置から得られた前記質問文から検索語集合と質問種別を判定する質問解析装置と、前記検索語集合と質問種別に従って、前記文書集合格納装置に格納された文書集合から関連文書集合を検索する文書検索装置と、関連文書集合中の各文書から質問文に対する回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として作成する回答抽出装置と、前記応答結果を出力する出力装置とを有するので、質問文に対する直接的な回答とその回答の元となる文書を提示することにより回答の信頼性をユーザが確認することができる。

【００１７】また、請求項６に記載の発明は、請求項５に記載の質問応答システムにおいて、前記文書検索装置は、関連文書集合の要素を文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索することを特徴とする。

【００１８】請求項６に記載の発明によれば、請求項５に記載の質問応答システムにおいて、前記文書検索装置は、関連文書集合の要素を文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索するので、直接的な回答が得られる。

【００１９】また、請求項７に記載の発明は、請求項５または６のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記文書検索装置が関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記回答抽出装置が前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの２つのスコアに従って、回答と文書の列を順序付けることを特徴とする。

【００２０】請求項７に記載の発明によれば、請求項５または６のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記文書検索装置が関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記回答抽出装置が前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの２つのスコアに従って、回答と文書の列を順序付けるようにしたので、質問文に対する回答の精度の向上が図れる。

【００２１】また、請求項８に記載の発明は、請求項５乃至７のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記関連文書集合の各文書から回答を抽出する際に、固有名詞や数値表現の認識を行なうことを特徴とする。

【００２２】請求項８に記載の発明によれば、請求項５乃至７のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記関連文書集合の各文書から回答を抽出する際に、固有名詞や数値表現の認識を行なうので、質問文に対する回答の精度の向上が図れる。

【００２３】また、請求項９に記載の発明は、文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問応答を行うための質問応答プログラムを記録したコンピュータ読み取り可能な記録媒体において、質問文を受け取る第１のステップと、入力された質問文から検索語集合と質問種別を判定する第２のステップと、前記検索語集合および質問種別に従って、前記与えられた文書集合から関連文書集合を検索する第３のステップと、前記関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を作成する第４のステップと、前記回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力するステップとをコンピュータに実行させる質問応答プログラムを記録した記録媒体を要旨とする。

【００２４】請求項９に記載の発明によれば、文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問応答を行うための質問応答プログラムを記録したコンピュータ読み取り可能な記録媒体において、質問文を受け取る第１のステップと、入力された質問文から検索語集合と質問種別を判定する第２のステップと、前記検索語集合および質問種別に従って、前記与えられた文書集合から関連文書集合を検索する第３のステップと、前記関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を作成する第４のステップと、前記回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力する第５のステップとをコンピュータに実行させる質問応答プログラムを記録媒体に記録したので、この記録媒体に記録した質問応答プログラムをコンピュータシステムに読み込ませ、実行することにより、質問文に対する直接的な回答とその回答の元となる文書を提示することにより回答の信頼性をユーザが確認することができる。

【００２５】

【発明の実施の形態】以下、本発明の実施の形態を、図面を参照して詳細に説明する。図１に本発明の実施の形態に係る質問応答システムの構成を示す。本発明の実施の形態に係る質問応答システムは、質問文に対する直接的な回答として文書の一部を取り出すとともに、回答を取り出した文書をユーザに出力することにより、回答の信頼性をユーザが確認できるようにしている。

【００２６】また、固有名詞の抽出技術や、与えられた文書集合から関連文書を検索する際における文書スコアおよび関連文書集合から回答を抽出する際における抽出スコアの両方を総合的に用いて評価することより、出力する回答と文書の列の順序を決めることにより、質問に対する正しい回答を回答と文書の列のより上位に並べることを可能にしている。

【００２７】すなわち、本発明の実施の形態に係る質問応答システムは、文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力するコンピュータシステムを使用した質問応答方法であって、入力された質問文から検索語集合と質問種別を判定し、該検索語集合および該質問種別に従って、前記与えられた文書集合から関連文書集合を検索し、該関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力することを特徴とする質問応答方法を実施するための装置である。

【００２８】図１において、本発明に係る質問応答システム１０は、与えられた文書集合を格納する文書集合格納装置１００と、質問文を受け取る入力装置１０２と、質問文を解析し、上記質問文から検索語集合と質問種別を判定する質問解析装置１０４と、検索語集合と質問種別に従って文書集合格納装置１００に格納されている文書集合から関連文書集合を検索する文書検索装置１０６と、関連文書集合から上記質問文に対する回答を抽出し、該回答と該回答を通出した文書の列を上記質問文に対する応答結果として作成する回答抽出装置１０８と、上記応答結果を出力する出力装置１１０とを有している。

【００２９】上記構成からなる本実施の形態に係る質問応答システム１０の処理内容を図２に示すフローチャートに基づいて説明する。まず、文書集合格納装置１００に、与えられた文書の集合が格納される（ステップ２００）。入力装置１０２は、質問文が入力されると、その質問文を質問解析装置１０４に渡す（ステップ２０１）。質問解析装置１０４は質問文に対して形態素解析を行い、質問文から検索語集合と質問種別を判定し、検索語集合と質問種別を文書検索装置１０６に渡す（ステップ２０２）。

【００３０】文書検索装置１０６は質問解析装置１０４から受け取った検索語集合と質問種別に従って、文書集合格納装置２に格納された文書集合を検索し、関連文書集合を抽出し、該関連文書集合の各文書に対して検索語が含まれている度合いを示す文書スコアを算出し、上記関連文書集合を、検索語集合及び質問種別と共に回答抽出装置１０８へ渡す（ステップ２０３）。

【００３１】回答抽出装置１０８は文書検索装置１０６から受け取った関連文書集合中の各文書について形態素解析を行う（ステップ２０４）。そして回答抽出装置１０８は、形態素解析を行った各文書に対して質問種別に従った単語を抽出対象とし、抽出すると共に（ステップ２０５）、上記抽出対象についてその抽出対象が含まれる文書中における検索語との距離に基づいて抽出スコアを算出する（ステップ２０６）。

【００３２】次いで、回答抽出装置１０８は、文書検索装置１０６で算出した文書スコアと上記抽出スコアに従って、上記関連文書集合中の各文書から質問文に対する回答を抽出し、この回答と、この回答を抽出した元となる文書の列を取り出し、回答及び文書の列の順序付けを行う（２０７）。そして回答抽出装置１０８は、順序付けを行った回答及び文書を示す文書番号の列を応答結果として出力装置１１０へ渡す。出力装置１１０は回答とその文書番号の列を応答結果として出力する（ステップ２０８）。

【００３３】なお、本実施の形態に係る質問応答システムを構成する各装置は論理的につながっていればよく、各装置が１台のコンピュータの上で通信しながら動く独立したプロセスとして実装されていても良いし、ネットワークで接続された複数のコンピュータに分散されて実装されていても良い。

【００３４】次に、本発明の実施の形態に係る質問応答システムの具体的動作について説明する。以下では、例として、質問文Ｑ１「アメリカの大統領は誰ですか？」に対して、回答「クリントン」と記事番号「Ｄ１」を得る例を述べる。まず、文書集合が表１に示すように、３つの文書Ｄ１、Ｄ２、Ｄ３からなるとする。

【表１】

文書集合

文書番号	文書
D1	「日本の森首相とクリントン・アメリカ大統領が会談した」
D2	「ロシアのプーチン大統領が来日した」
D3	「昨日,京都で祇園祭が行なわれた」

本発明の実施の形態では、説明を容易にするため、文書を１文だけで記述しているが、複数の文章からなる文書でもよい。また、質問文を１つ受けて、回答とその文書の列を返す例を述べているが、これを繰り返すことにより、質問と回答を繰り返すことができる。

【００３５】まず、文書集合格納装置１００に文書Ｄ１、Ｄ２、Ｄ３を格納する。ここでは、表形式で表現しているが、格納の方法は文書が格納できればリストやデータベースなどの他の方法であっても何でもよい。以下、表形式でデータを表した時には同様のことが言える。

【００３６】入力装置１０２は、質問文を受けとり、質問解析装置１０４へ渡す。まず、質問解析装置１０４は、質問文の質問種別を判定する。本実施の形態では、質問種別は人名を聞くｗｈｏ，場所を聞くｗｈｅｒｅ，日時を聞くｗｈｅｎの３種類とする。なお、この他の物の名前を聞くｗｈａｔや、方法を聞くｈｏｗなどの質問種別があったとしても、同様な方法で質問種別を判定することができる。

【００３７】質問種別の判定には、表２に示すような判定表現を持つ判定表を用いる。なお、判定表現は質問の種別を分けるために使える語であれば、他の語でもよい。

【表２】

質問種別の判定表現

質問種別	判定表現
who	だれ,誰,どの人
where	どこで,何処,どの国,国はどこ,何県
when	いつ,何時,何日,何月,何年,何曜日

判定表現が質問文Ｑ１に含まれるかどうかをチェックし、判定表現が含まれる質問種別をＱ１の質問種別ＱＴとする。Ｑ１には「誰」が含まれるので、Ｑ１の質問種別はｗｈｏとなる。

【００３８】質問解析装置１０４は、質問文を形態素解析し、単語に分けるとともに、品詞の情報を得る。形態素解析の手法は例えば文献（長尾真編：自然言語処理、岩波書店、１９９６）に述べられている。形態素解析は、辞書に含まれる単語が文に現れるかどうかを調べ、文を辞書にある単語の列に分割し、辞書に書かれている各単語の品詞のうち、前後の単語の関係から最適な品詞を選択することにより実現される。

【００３９】現在の例では、「アメリカの大統領は誰ですか？」は、表３のように形態素解析されるとする。

【表３】

Q1の形態素解析結果

単語番号	単語	品詞
1	アメリカ	固有名詞
2	の	格助詞
3	大統領	普通名詞
4	は	副助詞
5	誰	代名詞
6	です	助動詞
7	か	終助詞
8	?	記号

KW=｛アメリカ、大統領｝

と書ける。ここで、自立語のみを検索語集合とするのは、説明の簡単化と検索精度の向上を図るためであるので、すべての単語を検索語集合にしたり、他の選択法によって検索語集合を選択したりしてもかまわない。

【００４１】質問解析装置１０４は検索語集合ＫＷと質問種別ＱＴを文書検索装置１０６に渡す。文書検索装置１０６は、検索語集合ＫＷがより多く含まれる文書を文書集合格納装置１００に格納されている文書集合から探す。検索語集合の含まれている数を数え、それを各文書（文書番号で表される。）の文書スコアとする。この結果を表４に示す。

【表４】

文書スコア

文書番号	文書スコア
D1	2
D2	1
D3	0

RD=｛D1，D2｝

となる。なお、文書スコアの計算法はＩＤＦ法やＴＦＩＤＦ法など、当業界において用いられる方法ならなんでもよい。ＩＤＦ法やＴＦＩＤＦ法の計算式や検索の高速化のためのインデックス作成法は、例えば、文献（徳永健伸：情報検索と言語処理、東京大学出版会、１９９９年）に述べられている。

【００４４】また、ここでは文書スコアの計算には、各文書中に検索語集合の要素が現れるかどうかを文書中のなお、ここで品詞名は言語学や自然言語処理において利用されるものであれば何でもよい。例えば、普通名詞を一般名詞と表現してもよい。このうち、普通名詞、固有名詞などの自立語を検索語集合ＫＷとする。

【００４０】上記質問文から検索語集合ＫＷは、

（１）

【００４２】文書番号Ｄ１の文書には「アメリカ」、「大統領」の２つの検索語集合の要素が含まれるので、文書番号Ｄ１の文書の文書スコアは２となる。また、文書番号Ｄ２の文書スコアは、検索語集合の要素である「大統領」しか含まないので文書スコアは１となる。さらに、文書番号Ｄ３の文書は、検索語集合の要素を含まないので文書スコアは０である。

【００４３】文書検索装置１０６は文書スコアが０より大きい文書を関連文書集合ＲＤとする。したがって、

（２）

文字で調べたが、文書を形態素解析し、単語に分けてから文書中の単語と検索語を比較してもよい。さらに、文書全体を１まとまりとしなくても、文書の各段落を検索の単位としたり、１００文字といったパッセージを検索の単位としてもよい。文書検索装置１０６は関連文書集合ＲＤを、検索語集合及び質問種別と共に、回答抽出装置１０８に渡す。なお、関連文書集合ＲＤは文書全体を渡しても良いし、文書の名前だけを渡してもよい。

【００４５】回答抽出装置１０８は、関連文書集合ＲＤ中の文書を形態素解析する。文書番号Ｄ１と文書番号Ｄ２の各文書の形態素解析結果をそれぞれ表５、表６に示す。

【００４６】次に固有表現抽出法により、文書中の各単語が＜人名＞、＜地名＞、＜日時＞という３種類の固有表現のどれかに該当するかどうかを判定し、該当する場合はその種別を単語に付与する。基本的には固有名詞の辞書と周囲の単語により人名、地名、日時であるかを判定する。固有表現の抽出法については、例えば特開平１１－０６７５６２に記載されている。固有表現の判定後の結果を表７、表８に示す。

【００４７】

【表５】

文書D1の形態素解析結果

単語番号	単語	品詞
1	日本	固有名詞
2	の	格助詞
3	森	固有名詞
4	首相	普通名詞
5	と	副助詞
6	クリントン	固有名詞
7	・	記号
8	アメリカ	固有名詞
9	大統領	普通名詞
10	が	格助詞
11	会談	動詞
12	した	助動詞

【表６】

文書D2の形態素解析結果

単語番号	単語	品詞
1	ロシア	固有名詞
2	の	格助詞
3	プーチン	固有名詞
4	大統領	普通名詞
5	が	格助詞
6	来日	動詞
7	した	助動詞

【００４８】回答抽出装置１０８は、質問種別に従って単語を抽出対象とする。ｗｈｏの場合は品詞・固有表現が＜人名＞である単語を抽出対象とし、ｗｈｅｒｅの場合は品詞・固有表現が＜場所＞である単語を抽出対象とし、ｗｈｅｎの場合は品詞・固有表現が＜日時＞である単語を抽出対象とする。但し、質問種別に対応する抽出対象はこれに限るわけではない。例えば、ｗｈｅｒｅの抽出対象に組織名を加えても良い。また、“「我輩は猫である」”のようなカギ括弧で括られた部分を抽出対象に加えても良い。

【００４９】

【表７】

文書D1の固有表現抽出結果

単語番号	単語	品詞・固有表現
1	日本	＜地名＞
2	の	格助詞
3	森	＜人名＞
4	首相	普通名詞
5	と	副助詞
6	クリントン	＜人名＞
7	・	記号
8	アメリカ	＜地名＞
9	大統領	普通名詞
10	が	格助詞
11	会談	動詞
12	した	助動詞

【表８】

文書D2の固有表現抽出結果

単語番号	単語	品詞・固有表現
1	ロシア	＜地名＞
2	の	格助詞
3	プーチン	＜人名＞
4	大統領	普通名詞
5	が	格助詞
6	来日	動詞
7	した	助動詞

【００５０】本実施の形態の場合は、質問種別がｗｈｏであるので、文書番号Ｄ１の文書における単語「森」、「クリントン」、文書番号Ｄ２の文書における単語「プーチン」が抽出対象となる。これらの抽出対象について、検索語ＫＷとの距離を使って抽出スコアを計算する。ここでは、抽出対象と検索語ＫＷの各要素が何単語離れて出現するかを求めそれを距離とし、各要素について求められた距離の逆数の和を抽出スコアとする。

【００５１】具体的には、単語間の距離は２つの単語の単語番号の差の絶対値とする。文書番号Ｄ１の文書における単語「森」は単語「アメリカ」と５単語、単語「大統領」と６単語離れているので単語「森」の抽出スコアは１／５＋１／６＝0.37となる。また、文書番号Ｄ１の文書「クリントン」は「アメリカ」と２単語、「大統領」と３単語離れているので「クリントン」の抽出スコアは１／２＋１／３＝0.83となる。さらに、文書番号Ｄ１の文書における単語「プーチン」は単語「大統領」と１単語離れているので単語「プーチン」の抽出スコアは

１となる。なお、抽出スコアの計算は当業界で使われるものであれば、他のものでもよい。

【００５２】次に、回答抽出装置１０８は、各文書中の抽出対象について、総合スコアを次のような総合計算式で求める。

$$\text{総合スコア} = a \times \text{文書スコア} + b \times \text{抽出スコア} \quad (3)$$

但し、ａ、ｂはそれぞれ文書スコア、抽出スコアの重みを表すパラメータであり様々な定め方がある。ここでは、ａ＝１，ｂ＝１とする。なお、スコアの計算法は当業界で用いられるものであれば他の方法でも構わない。

【００５３】例えば、ａ、ｂ、ｃ、ｄをパラメータとして、

$$\text{総合スコア} = a \times \text{文書スコア} + b \times \text{段落スコア} + c \times \text{文スコア} + d \times \text{抽出スコア} \quad (4)$$

のように多段化してもよい。ここで、文書スコアの「文書」は、読点で区切られた複数の文容の集合体であり、文スコアの「文」は、読点までをひとかたまりとした文字列の集合体である。各文書中の抽出対象についての抽出スコア、文書スコア、総合スコアを計算した結果を表９に示す。

【００５４】

【表９】

抽出対象のスコア

抽出対象	文書	抽出スコア	文書スコア	総合スコア
森	D1	0.37	2	2.37
クリントン	D1	0.83	2	2.83
プーチン	D2	1	1	2

【００５５】さらに、回答抽出装置１０８は、抽出対象の総合スコアの大きい順に抽出対象を回答とし、その文書番号を出力装置１１０に渡し、結果を出力する。ここでは、次ように出力される。

１．回答＝「クリントン」、文書＝Ｄ１
２．回答＝「森」、文書＝Ｄ１
３．回答＝「プーチン」、文書＝Ｄ２

なお、出力の形式は同じ内容を表していれば、他の形式でもよい。このようにして、文書集合と「アメリカ大統領は誰ですか？」という入力文が与えられると、それに対する答えを文書集合から取り出し、回答とその回答を含む文書の列として出力された。

【００５６】尚、図２に示す処理内容を質問応答プログラムとして作成し、このプログラムをコンピュータにより読み取り可能な記録媒体に記録し、この記録媒体を使用することによりコンピュータシステムに質問応答プログラムを実行させることにより質問応答システムの機能を実現するようにしてもよい。

【００５７】すなわち、文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問応答を行うための質問応答プログラムを記録したコンピュータ読み取り可能な記録媒体において、質問文を受け取る第１のステップと、入力された質問文から検索語集合と質問種別を判定する第２のステップと、前記検索語集合および質問種別に従って、前記与えられた文書集合から関連文書集合を検索する第３のステップと、前記関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を作成する第４のステップと、前記回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力する第５のステップとをコンピュータに実行させる質問応答プログラムを記録媒体に記録し、この記録媒体に記録した質問応答プログラムをコンピュータシステムに読み込ませ、実行することにより、質問応答システムの機能を実現するようにしてもよい。

【００５８】なお、ここでいう「コンピュータシステム」とは、ＯＳや周辺機器等のハードウェアを含むものとする。また、「コンピュータ読み取り可能な記録媒体」とは、フロッピー（登録商標）ディスク、光磁気ディスク、ＲＯＭ、ＣＤ－ＲＯＭ等の可般媒体、コンピュータシステムに内蔵されるハードディスク等の記憶装置のことをいう。

【００５９】さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線等の通信回線を介してプログラムを送信する場合の通信線のように、短時間の間、動的にプログラムを保持するもの（伝送媒体ないしは伝送波）、その場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリのように、一定時間プログラムを保持しているものも含むものとする。

【００６０】また上記プログラムは、前述した機能の一部を実現するためのものであっても良く、さらに前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるもの、所謂差分ファイル（差分プログラム）であっても良い。

【００６１】以上説明したように、本発明の実施の形態に係る質問応答方法、質問応答システム及び質問応答プログラムを記録した記録媒体によれば、自然言語で表現された質問文に対する回答を求めることが可能となる。また、本発明の実施の形態に係る質問応答方法、質問応答システム及び質問応答プログラムを記録した記録媒体によれば、質問文に対する直接的な回答とその回答のも

とになる文書が求められるので、ユーザが直接回答を得るとともに、その回答の根拠を文書でチェックできる。

【００６２】従来の情報検索技術では、文書全体の列が提示されるため、上記実施の形態において質問文の内容である「大統領の名前」は、ユーザが文書を読んで初めて見つけることができたにすぎない。また、従来の情報抽出技術は分野に依存していたため、自由な質問文に対する答えを抽出することはできなかった。従来の質問応答システムは、回答とともに文書を提示していなかったので、ユーザが回答の正しさを計ることができなかった。

【００６３】さらに、従来の質問応答システムでは、日本語の処理に必要な単語分けを含む形態素解析を行なっていなかったり、固有表現の抽出を行なっていなかったり、抽出対象のスコアを評価する際に検索時のスコアを利用したりしていなかったため、回答の精度が低くなっていた。

【００６４】例えば、実施例において、検索スコアを無視して、総合スコアを抽出スコアとすると、回答の列の順序が、「プーチン」「クリントン」「森」の順になってしまう。また、形態素解析や固有表現の抽出を行なわないと、抽出対象の単語が決まらず、回答の精度が低くなる。本発明の実施の形態に係る質問応答方法、質問応答システム及び質問応答プログラムを記録した記録媒体は、このような問題を解決し、質問文に対する回答とその文書の列をユーザに提示できる効果がある。

【００６５】

【発明の効果】以上に説明したように、請求項１に記載の発明によれば、入力された質問文から検索語集合と質問種別を判定し、該検索語集合および該質問種別に従って、前記与えられた文書集合から関連文書集合を検索し、該関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力するようにしたので、質問文に対する直接的な回答とその回答の元となる文書を提示することにより回答の信頼性をユーザが確認することができる。

【００６６】請求項２に記載の発明によれば、請求項１に記載の質問応答方法において、前記関連文書集合の要素は文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索するようにしたので、直接的な回答が得られる。

【００６７】請求項３に記載の発明によれば、請求項１または２のいずれかに記載の質問応答方法において、前記関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの２つのスコアに従って、回答と文書の列を順序付けるようにしたので、質問文に対する回答の精度の向上が図れる。

【００６８】請求項４に記載の発明によれば、請求項１乃至３のいずれかに記載の質問応答方法において、前記関連文書集の各文書から回答を抽出する際に、固有名詞や数値表現の認識を行なうようにしたので、質問文に対する精度の向上が図れる。

【００６９】請求項５に記載の発明によれば、文書集合を格納する文書集合格納装置と、質問文を受けとる入力装置と、該入力装置から得られた前記質問文から検索語集合と質問種別を判定する質問解析装置と、前記検索語集合と質問種別に従って、前記文書集合格納装置に格納された文書集合から関連文書集合を検索する文書検索装置と、関連文書集合中の各文書から質問文に対する回答を抽出し、該回答と該回答を抽出した文書の列を前記質問文に対する応答結果として作成する回答抽出装置と、前記応答結果を出力する出力装置とを有するので、質問文に対する直接的な回答とその回答の元となる文書を提示することにより回答の信頼性をユーザが確認することができる。

【００７０】請求項６に記載の発明によれば、請求項５に記載の質問応答システムにおいて、前記文書検索装置は、関連文書集合の要素を文書全体ではなく、文書の一部として前記文書集合から前記関連文書集合を検索するので、直接的な回答が得られる。

【００７１】請求項７に記載の発明によれば、請求項５または６のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記文書検索装置が関連文書集合を検索する際に計算した各文書のスコアである文書スコアと前記回答抽出装置が前記関連文書集合の各文書から回答を抽出する際に計算した抽出スコアの２つのスコアに従って、回答と文書の列を順序付けるようにしたので、質問文に対する回答の精度の向上が図れる。

【００７２】請求項８に記載の発明によれば、請求項５乃至７のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記関連文書集合の各文書から回答を抽出する際に、固有名詞や数値表現の認識を行なうので、質問文に対する回答の精度の向上が図れる。請求項８に記載の発明によれば、請求項５乃至７のいずれかに記載の質問応答システムにおいて、前記回答抽出装置は、前記関連文書集合の各文書から回答を抽出する際に、固有名詞や数値表現の認識を行なうので、質問文に対する回答の精度の向上が図れる。

【００７３】請求項９に記載の発明によれば、文書集合と質問文が与えられると、該質問文に対する回答と文書の列を出力する質問応答を行うための質問応答プログラムを記録したコンピュータ読み取り可能な記録媒体において、質問文を受け取る第１のステップと、入力された質問文から検索語集合と質問種別を判定する第２のステップと、前記検索語集合および質問種別に従って、前記与えられた文書集合から関連文書集合の各文書から回答を抽出し、該回答と該回答を抽出した文書の列を作成する第４のステップと、前記回答と該回答を抽出した文書の列を前記質問文に対する応答結果として出力する第５のステップとをコンピュータに実行させる質問応答プログラムを記録媒体に記録したので、この記録媒体に記録した質問応答プログラムをコンピュータシステムに読み込ませ、実行することにより、質問文に対する直接的な回答とその回答の元となる文書を提示することにより回答の信頼性をユーザが確認することができる。

【図面の簡単な説明】

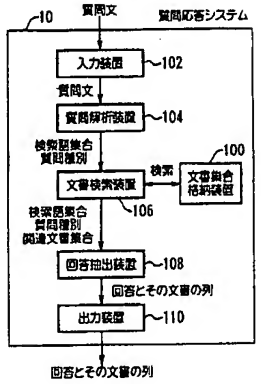
【図１】　本発明の実施の形態に係る質問応答システムの構成を示すブロック図。

【図２】　図１に示す質問応答システムの処理内容を示すフローチャート。

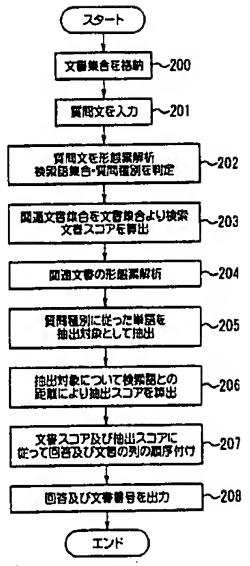
【符号の説明】

１０　質問応答システム
１００　文書集合格納装置
１０２　入力装置
１０４　質問解析装置
１０６　文書検索装置
１０８　回答抽出装置
１１０　出力装置

【図１】



【図２】



フロントページの続き

(72)発明者　平　博順
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

(72)発明者　賀沢　秀人
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

(72)発明者　廣田　啓一
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

(72)発明者　中島　浩之
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

(72)発明者　平尾　努
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

(72)発明者　加藤　恒昭
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

Ｆターム(参考)　5B075 ND03 NK32 PP24 PQ02 PQ74
5B091 AA11 AB06 CA02